Machine Learning 1

Lecture 4.5 - Supervised Learning
Bayesian Linear Regression - Predictive Distribution

Erik Bekkers

(Bishop 3.3.2)

Slide credits: Rianne van den Berg

Predictive Distribution

- Observed dataset with inputs $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N)^T$ and targets $\mathbf{t} = (t_1, t_2, ..., t_N)^T$

- Posterior distribution

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbb{1} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- Likelihood $p(t'|\mathbf{x}', \mathbf{w}, \beta) = \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x}')^T \mathbf{w}, \beta^{-1})$

- Predictive distribution for new input $\mathbf{x}'$

$$p(t'|\mathbf{x}', \mathbf{X}, \mathbf{t}, \alpha, \beta) = \int p(t' | \phi(\underline{x}')\underline{w}, \beta)\, p(\underline{w} | X, \underline{t}, \alpha, \beta)\, d\underline{w}$$

$$= \mathcal{N}(t' | \underline{\phi}(\underline{x}')^T \underline{m}_N, \sigma_N^2(\underline{x}'))$$

Bishop Eq. 2.115

$$\sigma_N^2(\mathbf{x}') = \frac{1}{\beta} + \underline{\phi}(\underline{x}')^T S_N \underline{\phi}(\underline{x}')$$

$$\lim_{N\to\infty} \sigma_N^2(\mathbf{x}') = \frac{1}{\beta}$$

Predictive Distribution

- Datasets generated with

$$t = \sin(2\pi x) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \beta^{-1})$$

$y(x,w)$

- Dataset sizes N = 1, 2, 4, 25

- Model:

$$y(x, \mathbf{w}) = \boldsymbol{\phi}(x)^T \mathbf{w}$$

- $\phi_j(x)$: Gaussian basis function

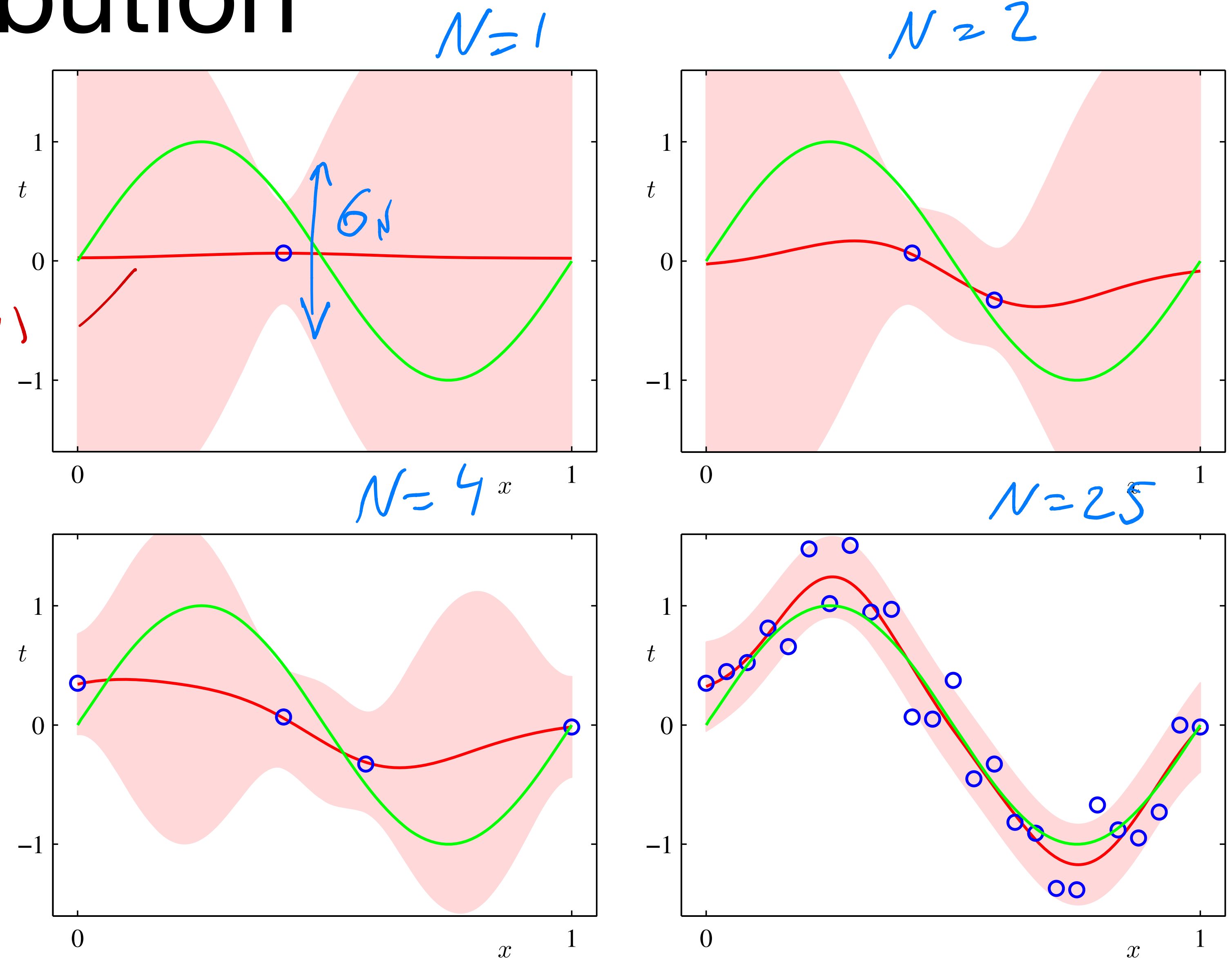

Figure: Predictive distribution (Bishop 3.8)

"Bayesian averaged weight"

$$p(t'|x', \mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\boldsymbol{\phi}(x')^T \mathbf{m}_N, \sigma_N^2(x'))$$

$$\sigma_N^2(x') = \frac{1}{\beta} + \boldsymbol{\phi}(x')^T \mathbf{S}_N \boldsymbol{\phi}(x')$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t} \qquad \mathbf{S}_N^{-1} = \alpha \mathbb{1} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}$$

① uncertainty is small near data points

② uncertainty decreases with larger N

Samples drawn from Bayesian Predictive Distribution

$w \sim$ $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ $\quad \mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$ $\quad \mathbf{S}_N^{-1} = \alpha \mathbb{1} + \beta \mathbf{\Phi}^T \mathbf{\Phi}$

$y(x, w)$

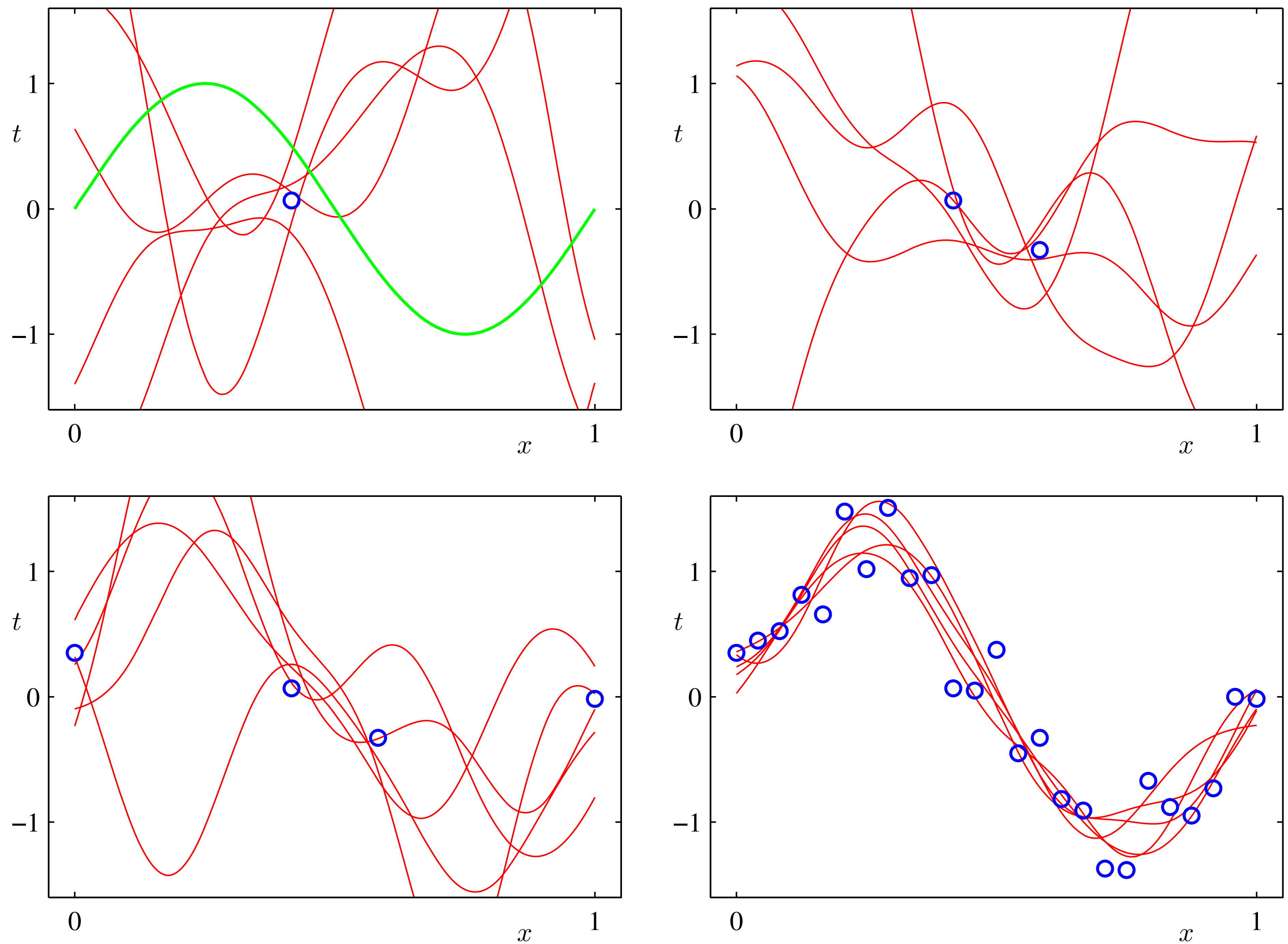

Figure: Sample functions y(x,w) with w sampled from posterior distribution (Bishop 3.9)